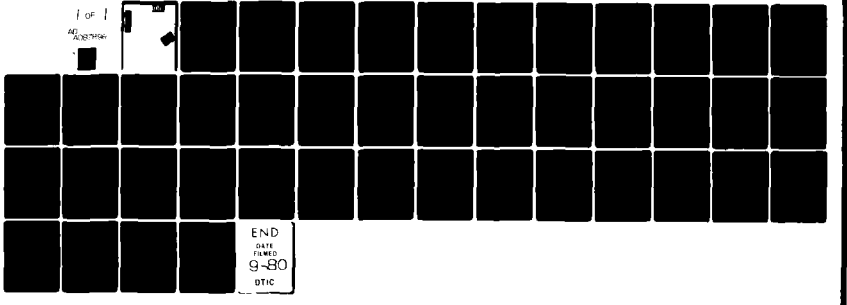
AD-A087 898 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
DETECTING CHANGE POINTS IN SAMPLING FROM MULTINOMIAL DISTRIBUTI--ETC(U)
MAR 80 A E GELFAND N00014-76-C-0475
UNCLASSIFIED TR-282 NL



LEVEL

AD A087898

(11)

DETECTING CHANGE POINTS IN SAMPLING FROM MULTINOMIAL DISTRIBUTIONS

By

ALAN E. GELFAND

TECHNICAL REPORT NO. 282

March 4, 1980

DTIC
ELECTE
AUG 15 1980
S C D

Prepared under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted
for any Purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Detecting Change Points in Sampling from Multinomial Distributions

Alan E. Gelfand

1. Introduction

Suppose that we are observing a sequence of categorical variables. Our problem addresses the matter of if and when a change in the underlying cell probabilities has occurred. In other words, we are trying to detect, along the sequence a shift from one multinomial distribution to another.

The literature to date has discussed the case of shifts for one-dimensional observations (i.e., binomial shifts in our setting). Page [3] studies a cumulative sum over time. Chernoff and Zacks [1] showed that a particular weighted sum weighing recent observations more heavily arises from a Bayesian approach. Kander and Zacks [2] generalized this result to random variables whose distributions belong to the one parameter exponential family. The latter two articles develop the problem from a hypothesis testing point of view. Sclove [4] examines these results in the binomial case. We shall see that the multidimensional problem is a bit less tractable. Sclove also suggests two possible applications. One application is in quality

control where one might wish to detect whether a production process has shifted from being "in control" to being "out of control" (i.e., from a low probability of producing a defective item to a higher one). A second application provides an epidemiology context where we would be concerned with whether the probability of contracting a disease has changed.

A broader application is the detection problem. As a given field is scanned there is a certain probability of detecting an object therein. If at some unknown time point something occurs in the field resulting in a change in the probability of detection then we are precisely in the situation described. The first example is easily extended to a multinomial framework if the production process yields items manufactured according to specification limits. We then have three natural categories: (i) the item is below the lower specification limit, (ii) the item is within the specification limits and (iii) the item is above the upper specification limit. The third example is easily extended by further elaborating the detection and no detection classifications. Yet another example involves attempting to detect shifts in political, sociological or psychological response over time with respect to categorical questions.

In certain applications the initial cell probabilities may be known, i.e., the probability of a defective

when the process is in control or the probability of contracting a disease under nonepidemic conditions. In other applications no probabilities will be known. We confine ourselves to this latter circumstance.

In all of our examples it may be that change occurs gradually over time. However we will presume that each change is somewhat sharp and that the changed distribution persists for a long period of time relative to the frequency of observation before a next change occurs. We consider exclusively the probability of detecting the first distributional change. If several distributional shifts may be expected then subsequent changes would be discovered by continuous monitoring of the process using the procedures developed in the subsequent sections.

The format of the paper, then, is the following. In section 2 we formalize the problem and developed weighted and unweighted cumulative statistics and their properties. In the third section we create one-dimensional decision functions of these statistics according to three different motivations which lead to the most effective procedure we have obtained thus far. Finally in section 4 we examine a portion of the results of a large simulation study.

2. Preliminaries

Formally our problem is the following. Vector valued observations X_i are taken with components X_{ij},

$j=1,\ldots,r$ such that $\underset{\sim}{X}_i$ is distributed as a generalized binomial random variable, i.e., one multinomial trial with associated probability vector $\underset{\sim}{p}$ with components $p_1,p_2,\ldots,p_r$ for $i=1,\ldots,k$ while X_i is distributed as a generalized binomial random variable with associated probability vector $\underset{\sim}{p}'$ with components $p_1',\ldots,p_r'$ for $i=k+1,k+2,\ldots$.

The vectors $\underset{\sim}{p}$ and $\underset{\sim}{p}'$ are assumed unknown along with the change point k which is to be estimated.

At any given trial ℓ let us define the following statistics

$$(1) \qquad S_{m,j}^{(\ell)} = \sum_{i=\ell-m+1}^{\ell} X_{ij}$$

$$j=1,2,\ldots,r$$

$$T_{m,j}^{(\ell)} = \sum_{i=\ell-m+1}^{\ell-1} (m+i-\ell)X_{i+1,j}$$

Let $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$ be rx1 vectors whose components are the $S_{m,j}^{(\ell)}$ and $T_{m,j}^{(\ell)}$ respectively.

The S statistic, for a given j, is just an unweighted sum of the last m X_{ij}'s up to and including $X_{\ell j}$. The T statistic for a given j is a weighted sum of the last m X_{ij}'s with greater weight attached to the latter observations. Over increasing ℓ $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$ will be referred to an unweighted and weighted moving sums respectively.

The $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$ may be examined directly over ℓ to uncover evidence of a distributional change. If for a given j $p_j < p_j'$ then $S_{m,j}^{(\ell)}$ ought to increase after the change point and $T_{m,j}^{(\ell)}$ even more so. A similar statement holds when $p_j > p_j'$. In fact such examination of the $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$ may be helpful in confirming shifts suggested by the approaches presented in section 3. Of course, by themselves they hardly constitute a precisely defined procedure.

However the intuition incorporated into these statistics argues that we should examine the $\underset{\sim}{X}_i$'s in blocks sequentially and likely with overlap. Of course to be able to effectively detect a change point we must assume that k is large with respect to the block size m. In selecting m we face a trade off. Use of a large m reduces noise, i.e., achieves stability of our $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$ vectors under no distributional shift while use of a small m makes the moving average more responsive to the incidence of such a shift. In addition, the choice of m must depend (in an obviously monotonic increasing fashion) on the known number of categories, r. If m is too small relative to r, many of the cell frequencies will be zero, i.e., many of the components of $S_m^{(\ell)}$ and $T_m^{(\ell)}$ will be zero regardless of whether or not a change has occurred. The only shifts we could hope to detect would be in the most probable categories thereby essentially degenerating the problem to a binomial case.

How do we justify the use of $\underset{\sim}{S}_m^{(\ell)}$ and $\underset{\sim}{T}_m^{(\ell)}$. The $S_{m,j}^{(\ell)}$ arise rather naturally. They are the raw cell frequencies. Cumulation is suggested by a variety of asymptotic considerations. The $T_{m,j}$ are less intuitive. A more general weighted moving sum would be of the form

$$\sum_{i=\ell-m+1}^{\ell-1} c(m,\ell,i)X_{i+1,j}.$$

The selection of $c(m,\ell,i) = m+i-\ell$ develops as follows. For the one-dimensional problem in the exponential family Kander and Zacks showed that for testing a single change in the mean (2) is Bayes against a uniform prior on the time of change. Let us consider this approach in our multidimensional situation.

We suppress ℓ and consider a fixed sequence of m observations. We wish to test the hypothesis of no distributional change across the m observations against the alternative of a single change. We have independent $\underset{\sim}{X}_i \sim GB(\underset{\sim}{p}_i)$, $i=1,\ldots,m$ with

$$H_o: \ \underset{\sim}{p}_1 = \underset{\sim}{p}_2 = \cdots = \underset{\sim}{p}_m = \underset{\sim}{p}$$

and

$$H_A: \ \underset{\sim}{p}_1 = \underset{\sim}{p}_2 = \cdots = \underset{\sim}{p}_k = \underset{\sim}{p},$$

$$\underset{\sim}{p}_{k+1} = \underset{\sim}{p}_{k+2} = \cdots = \underset{\sim}{p}_m = \underset{\sim}{p}'.$$

Since k is unknown we will suppose k is random, distributed according to $\tau(k)$, $k=1,2,\ldots,m-1$. We will specialize τ to the discrete uniform later. The conditional distribution of the sample given k is

$$f(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m|k) = \prod_{i=1}^{k}\prod_{j=1}^{r} p_j^{x_{ij}} \prod_{i=k+1}^{m}\prod_{j=1}^{r} p_j^{x_{ij}}$$

$$= \prod_{j=1}^{r} p_j^{S_{k,j}} \prod_{j=1}^{r} p_j^{S_{m,j}-S_{k,j}}.$$

Thus the unconditional distribution of the sample is

$$(3)\quad f(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m) = \sum_{k=1}^{m-1}\prod_{j=1}^{r} p_j^{S_{k,j}} \prod_{j=1}^{r} p_j^{S_{m,j}-S_{k,j}}\tau(k).$$

Consider $g(\underset{\sim}{p})$ such that $g^{(j)}(\underset{\sim}{p}) = \frac{\partial g}{\partial p_j}$ exists on $[0,1]$, $j=1,\ldots,r$. Let $p_j' = p_j+\delta_j$. We thus have as $||\underset{\sim}{\delta}||\to 0$

$$g(\underset{\sim}{p}') \approx g(\underset{\sim}{p}) + \Sigma\, g^{(j)}(\underset{\sim}{p})\delta_j.$$

In particular with $g(p) = \Sigma\, c_j \log p_j$

$$(4)\qquad g(\underset{\sim}{p}') \approx \sum_{j=1}^{r} c_j \log p_j + \sum_{j=1}^{r} \frac{e_j\delta_j}{p_j}.$$

Rewriting (3) as

Accession For
NTIS GRA&I
DDC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or special
A

$$f(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)=\sum_{k=1}^{m-1}\tau(k)\cdot e^{\sum_{j=1}^{r} S_{k,j}\log p_j + \sum_{j=1}^{r}(S_{m,j}-S_{k,j})\log p_j'}$$

and using (4) on the last summation yields

$$f(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)\underset{\sim}{\sim}\sum_{k=1}^{m-1}\tau(k)e^{\sum_{j=1}^{r} S_{k,j}\log p_j+\sum_{j=1}^{r}(S_{m,j}-S_{,,j})\log p_j}$$

$$+\sum_{j=1}^{r}(S_{m,j}-S_{k,j})\delta_j/p_j$$

$$\underset{\sim}{\sim}\sum_{k=1}^{m-1}\tau(k)e^{\sum_{j=1}^{r} S_{m,j}\log p_j+\sum_{j=1}^{r}(S_{m,j}-S_{k,j})\delta_j/p_j}.$$

Under the null hypothesis

$$f(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m) = e^{\sum_{j=1}^{r} S_{m,j}\log p_j}$$

so that the likelihood ratio $\lambda(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m)$ becomes

$$(5)\quad \lambda(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m) = \sum_{k=1}^{m-1}\tau(k)e^{\sum_{j=1}^{r}(S_{m,j}-S_{k,j})\delta_j/p_j}.$$

Again as $||\underset{\sim}{\delta}||\to 0$ $e^{c_j\delta_j/p_j}\underset{\sim}{\sim}1+c_j\delta_j/p_j$ so that (5) becomes, ignoring terms involving $\delta_{j_1}\cdot\delta_{j_2}$

$$\lambda(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m) \approx \sum_{k=1}^{m-1} \tau(k) \sum_{j=1}^{r} (S_{m,j}-S_{k,j})\delta_j/p_j+1.$$

If we now let $\tau(k) = 1/(m-1)$ and interchange the order of summation we obtain

$$\lambda(\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m) = \frac{1}{m-1} \sum_{j=1}^{r} \frac{\delta_j}{p_j} \sum_{k=1}^{m-1} (S_{m,j}-S_{k,j})+1$$

$$= \frac{1}{m-1} \sum_{j=1}^{r} \frac{\delta_j}{p_j} T_{m,j}+1.$$

Thus a distributional shift would be indicated by large values of

$$\sum_{j=1}^{r} \frac{\delta_j}{p_j} T_{m,j}. \tag{6}$$

If $\underset{\sim}{p}$ and $\underset{\sim}{p}'$ were known (6) could be used directly and would provide the Bayes test statistic. With $\underset{\sim}{p}$ and $\underset{\sim}{p}'$ unknown, some δ_j will be positive, some negative and some zero but the given linear combination of the $T_{m,j}$ is not computable. The δ_j and p_j can be estimated from the sample. (We shall examine this point later.) However this would necessitate making $2(r-1)$ estimates from m observations and will provide hopelessly unstable estimators. Hence the most we can conclude at the moment is that the $T_{m,j}$ seem to be appropriate weighted averages to study over time but exactly how to combine them into a

one-dimensional test statistic remains to be discussed.

In concluding this section let us examine the behavior of $\underset{\sim}{S}_m$ and $\underset{\sim}{T}_m$. We will continue to suppress ℓ for the remainder of this section.

Note that $\sum_{j=1}^{r} S_{m,j} = m$ and thus as we proceed over time, some $S_{m,j}$ will increase while others decrease but their sum, at any fixed time is m. Moreover if we continue to assume a uniform distribution for the change point

$$
\begin{aligned}
E(S_{m,j}) &= E(E(S_{m,j}|k)) \\
&= \frac{1}{m-1} \sum_{k=1}^{m-1} E\left(\sum_{i=1}^{m} X_{ij} | k \right) \\
&= \frac{1}{m-1} \sum_{k=1}^{m-1} (kp_j + (m-k)p_j') \\
&= \frac{m(p_j + p_j')}{2}
\end{aligned}
\tag{7}
$$

which equals mp_j under no change.

$$
\begin{aligned}
\operatorname{var}(S_{m,j}) &= \operatorname{var}(E(S_{m,j}|k)) + E(\operatorname{var}(S_{m,j}|k)) \\
&= \operatorname{var}(kp_j + (m-k)p_j') + E\left(\operatorname{var}\left(\sum_{i=1}^{m} X_{ij} | k \right)\right)
\end{aligned}
$$

$$= (p_j - p_j')^2 \text{var}(k) + E(k\, p_j(1-p_j) + (m-k)p_j'(1-p_j'))$$

$$= (p_j - p_j')^2 \frac{m(m-2)}{12} + \frac{m}{2}(p_j(1-p_j) + p_j'(1-p_j'))$$

which equals $m\, p_j(1-p_j)$ under no change.

In fact a similar calculation on the individual observations shows that X_{ij} is a single Bernoulli trial with success probability

$$p_{ij} = \frac{(i-1)p_j' + (m-i)p_j}{m-1}. \tag{8}$$

Hence the exact distribution of $S_{m,j}$ is that of the sum of m independent but non-identically distributed Bernoulli random variables. Under the null hypothesis it is, of course, $Bi(m,p_j)$. Moreover from (8) the moments of X_{ij} can be readily computed although the expressions become rather unwieldy. In particular we can show that

$$\frac{[\sum_{i=1}^{m} E(X_{ij} - E(X_{ij}))^2]^{1/2}}{[\sum_{i=1}^{m} E(|X_{ij} - EX_{ij}|^3)]^{1/3}} = \bigcirc(m^{-2/3}).$$

Thus Liapunov's theorem insures that $S_{m,j}$ is asymptotically normal under either hypothesis.

Turning to $\underset{\sim}{T}_m$ we note that upon interchanging order of summation

$$\sum_{j=1}^{r} T_{m,j} = \frac{m(m-1)}{2}.$$

Thus as with $\underset{\sim}{S}_m$ as we proceed over time, some $T_{m,j}$ will increase while others will decrease but their sum at any fixed time is $\frac{m(m-1)}{2}$. Similarly we may show that

$$(9) \qquad E(T_{m,j}) = \frac{m(m-2)}{6} p_j + \frac{m(2m-1)}{6} p_j'$$

which becomes $\frac{m(m-1)}{2} p_j$ under no change. Furthermore

$$\text{var}(T_{m,j}) = \sum_{i=1}^{m} i^2 \text{var } X_{i+1,j} = \sum_{i=2}^{m} (i-1)^2 p_{ij}(1-p_{ij}).$$

This is a rather messy expression which obviously reduces to $\frac{m(m-1)(2m-1)}{6} p_j(1-p_j)$ under no change. Liapunov's theorem again insures that $T_{m,j}$ is asymptotically normal. Kander and Zacks develop this result for the more general exponential family case and also discuss the rate of convergence. They briefly consider the exact distribution of $T_{m,j}$ in a scaled binomial case (p. 1202).

The expressions in (7) and (9) enable us to construct method of moments type estimators of p_j, p_j' (and δ, if desired), i.e.,

$$\hat{p}_j = \frac{\frac{m(2m-1)}{6}S_{m,j} - \frac{m}{2}T_{m,j}}{\frac{m^2(m+1)}{12}} = \frac{2(2m-1)S_{m,j} - 6T_{m,j}}{m(m+1)}$$

$$\hat{p}'_j = \frac{\frac{m}{2}T_{m,j} - \frac{m(m-2)}{6}S_{m,j}}{\frac{m^2(m+1)}{12}} = \frac{6T_{m,j} - 2(m-2)S_{m,j}}{m(m+1)}$$

and

$$\hat{\delta}_j = \frac{12T_{m,j} - 6(m-1)S_{m,j}}{m(m+1)}.$$

All of these estimators are unbiased. However since

$$\mathrm{cov}(T_{m,j}, S_{m,j}) = \sum_{i=2}^{m} (i-1)\,\mathrm{var}(X_{ij}) = \sum_{i=2}^{m} (i-1)p_{ij}(1-p_{ij})$$

none of these estimators is consistent. All three have variance of order m. Thus effective estimation of the coefficients of the $T_{m,j}$ in (6) is revealed to be hopeless unless the change point is known.

3. The Methods

We now consider a variety of procedures which suggest themselves as plausible methods for detecting a distributional shift. These methods may be loosely classified under three headings -- (i) Law of Large Numbers type approaches, (ii) Departures from centrality type

approaches and (iii) Tests of Homogeneity type approaches. We defer comparisons and criticisms of these approaches to the next section.

3.1 "Law of Large Numbers" Approaches

Consider the absolute difference

$$|S_{m,j}^{(\ell)} - S_{m,j}^{(\ell-1)}| = |X_{\ell,j} - X_{\ell-m,j}|$$

and define

$$Q_m^{(\ell)} = \sum_{j=1}^{r} |S_{m,j}^{(\ell)} - S_{m,j}^{(\ell-1)}|. \tag{10}$$

Note that

$$Q_m(\ell) = \begin{cases} 0 & \text{if} \quad X_{\ell,j} = X_{\ell-m,j},\ j=1,\ldots,r \\ 2 & \text{otherwise} \end{cases}$$

and thus

$$P(Q_m(\ell) = 0) = \begin{cases} \Sigma p_j^{\,2} & \text{if} \quad \ell \le k \\ \Sigma p_j p_j' & \text{if} \quad k < \ell \le k+m \\ \Sigma p_j'^{\,2} & \text{if} \quad \ell > k+m \end{cases}$$

with $P(Q_m(\ell) = 2) = 1 - P(Q_m(\ell) = 0)$. Hence

$$E(Q_m(\ell)) = \begin{cases} 2(1 - \Sigma p_j^2) & \text{if} \quad \ell \leq k \\ 2(1 - \Sigma p_j p_j') & \text{if} \quad k < \ell \leq k+m \\ 2(1 - \Sigma p_j'^2) & \text{if} \quad \ell > k+m \end{cases}$$

$$\text{var}(Q_m(\ell)) = \begin{cases} 4\,\Sigma p_j^2(1 - \Sigma p_j^2) & \text{if} \quad \ell \leq k \\ 4\,\Sigma p_j p_j'(1 - \Sigma p_j p_j') & \text{if} \quad k < \ell \leq k+m \\ 4\,\Sigma p_j'^2(1 - \Sigma p_j'^2) & \text{if} \quad \ell > k+m. \end{cases}$$

Similarly, consider the absolute difference

$$\left| T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)} \right| = \left| \sum_{i=\ell-m+1}^{\ell-1} (m+i-\ell) X_{i+1,j} - \sum_{i=\ell-m}^{\ell-2} (m+i-\ell+1) X_{i+1,j} \right|$$

$$= \left| m\, X_{\ell,j} - S_{m,j}^{(\ell)} \right|.$$

Hence

$$\left| T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)} \right| = \begin{cases} S_{m,j}^{(\ell)} & \text{if} \quad X_{\ell,j}=0 \\ m - S_{m,j}^{(\ell)} & \text{if} \quad X_{\ell,j}=1 \end{cases}$$

and thus

$$P(|T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)}| = S_{m,j}^{(\ell)}) = \begin{cases} 1-p_j & \text{if} \quad \ell \leq k \\ 1-p_j' & \text{if} \quad \ell > k \end{cases}$$

and $P(|T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)}| = m - S_{m,j}^{(\ell)}) = 1 - P(|T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)}| = S_{m,j}^{(\ell)})$.

Define

$$R_m(\ell) = \sum_{j=1}^{r} |T_{m,j}^{(\ell)} - T_{m,j}^{(\ell-1)}|. \tag{11}$$

Then

$$R_m(\ell) = m - S_{m,j_o}^{(\ell)} + \sum_{\substack{j=1 \\ j \neq j_o}}^{r} S_{m,j}^{(\ell)}, \quad \text{if} \quad X_{\ell,j_o} = 1$$

$$= 2(m - S_{m,j_o}^{(\ell)}) \quad , \quad \text{if} \quad X_{\ell,j_o} = 1.$$

Hence $0 \leq R_m(\ell) \leq 2(m-1)$ since $S_{m,j_o}^{(\ell)} \geq X_{\ell,j_o} = 1$ and $R_m(\ell)$ takes on values $0,2,\ldots,2(m-1)$. Furthermore for $a = 0,1,2,\ldots,m-1$

$$P(R_m(\ell) = 2a) = \sum_{j=1}^{r} P(R_m^{(\ell)} = 2a, X_{\ell,j} = 1)$$

$$= \sum_{j=1}^{r} P(S_{m,j}^{(\ell)} = m-a, X_{\ell,j} = 1)$$

$$= \sum_{j=1}^{r} P(S_{m-1,j}^{(\ell-1)} = m-1-a) P(X_{\ell,j} = 1).$$

The distribution of $S^{(\ell-1)}_{m-1,j}$ was discussed in the previous section under a uniform distribution over the change point k. For the present k is fixed so that if $\ell \leq k+1$, $S^{(\ell-1)}_{m-1,j} \sim Bi(m-1,p_j)$. If $\ell \geq k+m$ $S^{(\ell-1)}_{m-1,j} \sim Bi(m-1,p_j')$. If $k+1<\ell<k+m$, $S^{(\ell-1)}_{m-1,j} = W_1 + W_2$ where W_1 and W_2 are independent with $W_1 \sim Bi(m+k-\ell,p_j)$ and $W_2 \sim Bi(\ell-1-k, p_j')$. Of course $P(X_{\ell,j}=1) = p_j$ if $\ell \leq k$ and $= p_j'$ if $\ell > k$. Let

$$n_1(\ell) = \begin{cases} m-1 & \ell \leq k+1 \\ m+k-\ell & k+1<\ell<k+m \\ 0 & \ell \geq k+m \end{cases}$$

$$n_2(\ell) = \begin{cases} 0 & \ell \leq k+1 \\ \ell-1-k & k+1<\ell<k+m \\ m-1 & \ell \geq k+m \end{cases}$$

and

$$\gamma(m,\ell,a,j) = \sum_{i=0}^{m-a-1} \binom{n_1(\ell)}{i} p_j^{\,i} (1-p_j)^{n_1(\ell)-i} \binom{n_2(\ell)}{m-a-1-i} p_j'^{\,m-a-1-i} (1-p_j')^{n_2(\ell)-(m-a-1-i)}$$

with $\binom{c}{d} = 0$ if $c<d$ and $\binom{0}{0} \equiv 1$. Then

$$P(R_m(\ell) = 2a) = \begin{cases} \sum_{j=1}^{r} \gamma(m,\ell,a,j)p_j & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r} \gamma(m,\ell,a,j)p_j' & \text{if} \quad \ell > k. \end{cases}$$

Despite the awkwardness of the distribution of $R_m(\ell)$ its mean and variance are not that difficult to compute. From the argument after expression (11)

$$E(R_m(\ell)) = \begin{cases} \sum_{j=1}^{r} E[2(m-S_{m,j}^{(\ell)})|X_{\ell,j}=1]p_j & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r} E[2(m-S_{m,j}^{(\ell)})|X_{\ell,j}=1]p_j' & \text{if} \quad \ell > k \end{cases}$$

$$= \begin{cases} \sum_{j=1}^{r} 2[m-1-E(S_{m-1,j}^{(\ell-1)}]p_j & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r} 2[m-1-E(S_{m-1,j}^{(\ell-1)})]p_j' & \text{if} \quad \ell > k \end{cases}$$

$$= \begin{cases} \sum_{j=1}^{r} 2(m-1)p_j(1-p_j) & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r} 2[m-1-((m+k-\ell)p_j+(\ell-1-k)p_j')]p_j' & \text{if} \quad k+1 \leq \ell < k+m \\ \sum_{j=1}^{r} 2(m-1)p_j'(1-p_j') & \text{if} \quad \ell \geq k+m \end{cases}$$

$$= \begin{cases} 2(m-1)(1-\Sigma p_j^2) & \text{if} \quad \ell \leq k \\ 2(m+k-\ell)(1-\Sigma p_j p_j') + 2(\ell-1-k)(1-\Sigma p_j'^2) & \\ \qquad \text{if} \quad k+1 \leq \ell \leq k+m & \\ 2(m-1)(1-\Sigma p_j'^2) & \text{if} \quad \ell > k+m. \end{cases}$$

Similarly we may obtain the variance of $R_m(\ell)$ by computing

$$E(R_m^2(\ell)) = \begin{cases} \sum_{j=1}^{r} p_j E[(2(m-S_{m,j}^{(\ell)}))^2 | X_{\ell,j}=1], & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r} p_j' E[(2(m-S_{m,j}^{(\ell)}))^2 | X_{\ell,j}=1], & \text{if} \quad \ell > k \end{cases}$$

$$= \begin{cases} 4\sum_{j=1}^{r} p_j E(m-1-S_{m-1,j}^{(\ell-1)})^2, & \text{if} \quad \ell \leq k \\ 4\sum_{j=1}^{r} p_j' E(m-1-S_{m-1,j}^{(\ell)})^2, & \text{if} \quad \ell > k \end{cases}$$

$$= \begin{cases} 4\sum_{j=1}^{r} p_j [\text{var}(S_{m-1,j}^{(\ell-1)}) + (m-1-E\, S_{m-1,j}^{(\ell-1)})^2] & \\ \qquad \text{if} \quad \ell \leq k & \\ 4\sum_{j=1}^{r} p_j' [\text{var}(S_{m-1,j}^{(\ell-1)}) + (m-1-E\, S_{m-1,j}^{(\ell-1)})^2] & \\ \qquad \text{if} \quad \ell > k. & \end{cases}$$

From previous discussion we know the mean and variance of $S_{m-1,j}^{(\ell-1)}$ so that

$$E(R_m^2(\ell)) = \begin{cases} 4\sum_{j=1}^{r}[(m-1)p_j^2(1-p_j)+(m-1)^2(1-p_j)^2p_j] \\ \qquad \text{if} \quad \ell \leq k \\ 4\sum_{j=1}^{r} p_j'[(m+k-\ell)p_j(1-p_j)+(\ell-1-k)p_j'(1-p_j') \\ \qquad +((m+k-\ell)(1-p_j)+(\ell-1-k)(1-p_j'))^2] \\ \qquad \text{if} \quad k+1 \leq \ell \leq k+m \\ 4\sum_{j=1}^{r}[(m-1)p_j'^2(1-p_j')+(m-1)^2(1-p_j')^2p_j'] \\ \qquad \text{if} \quad \ell > k+m. \end{cases}$$

Subtracting $(E(R_m(\ell)))^2$ and simplifying yields

$$\text{var}(R_m(\ell)) = \begin{cases} 4(m-1)\Sigma p_j^2(1-p_j)+4(m-1)^2(\Sigma p_j^2-(\Sigma p_j^2)^2) \\ \qquad \text{if} \quad \ell \leq k \\ 4(m+k-\ell)\Sigma p_j' p_j(1-p_j)+4(\ell-1-k)\Sigma p_j'^2 \\ \qquad +4(m+k-\ell)^2(\Sigma p_j' p_j^2-(\Sigma p_j p_j')^2) \\ \qquad +4(\ell-1-k)^2(\Sigma p_j'^3-(\Sigma p_j'^2)^2) \\ \qquad +4(m+k-\ell)(\ell-1-k)(\Sigma p_j p_j'^2-\Sigma p_j p_j' \Sigma p_j'^2) \\ \qquad \text{if} \quad k+1 \leq \ell \leq k+m \\ 4(m-1)\Sigma p_j'^2(1-p_j')+4(m-1)^2(\Sigma p_j'^3-(\Sigma p_j'^2)^2) \\ \qquad \text{if} \quad \ell > k+m. \end{cases}$$

How may we employ (10) and (11) to develop detection procedures. Note that

$$E(1 - \frac{Q_m(\ell)}{2}) = E(1 - \frac{R_m(\ell)}{2(m-1)}) = \Sigma p_j^2 \quad \text{if} \quad \ell \leq k$$

$$= \Sigma p_j'^2 \quad \text{if} \quad \ell > k+m$$

and

$$E(1 - \frac{Q_m(\ell)}{2}) = \Sigma p_j p_j' \quad \text{if} \quad k+1 \leq \ell \leq k+m$$

$$E(1 - \frac{R_m(\ell)}{2(m-1)}) = \frac{(m+k-\ell)\Sigma p_j p_j' + (\ell-1-k)\Sigma p_j'^2}{m-1}$$

$$\text{if} \quad k+1 \leq \ell \leq k+m.$$

It is also apparent that

$$\text{var}(1 - \frac{R_m(\ell)}{2(m-1)}) \leq \text{var}(1 - \frac{Q_m(\ell)}{2}) \quad \text{for } \ell < k \text{ and for } \ell > k+m$$

i.e., for $\ell \leq k$

$$[\Sigma p_j^2(1 - \Sigma p_j^2)] - [\frac{1}{m-1}\Sigma p_j^2(1-p_j) + \Sigma p_j^3 - (\Sigma p_j^2)^2]$$

$$= (\Sigma p_j^2 - \Sigma p_j^3)(\frac{m-2}{m-1}) \geq 0$$

with a similar argument for $\ell \geq k+m$.

Furthermore the Cauchy-Schwarz inequality assures that

$$\Sigma p_j p_j' \leq \sqrt{\Sigma p_j^2 \Sigma p_j'^2}$$

i.e., $$\Sigma p_j p_j' \leq \max(\Sigma p_j^2, \Sigma p_j'^2)$$

so that we must have one of the following

(i) $\Sigma p_j p_j' \leq \Sigma p_j^2 \leq \Sigma p_j'^2$

(ii) $\Sigma p_j p_j' \leq \Sigma p_j'^2 \leq \Sigma p_j^2$

(iii) $\Sigma p_j^2 \leq \Sigma p_j p_j' \leq \Sigma p_j'^2$

(iv) $\Sigma p_j'^2 \leq \Sigma p_j p_j' \leq \Sigma p_j^2$.

Thus if we were to monitor $1 - \frac{Q_m(\ell)}{2}$ or $1 - \frac{R_m(\ell)}{2(m-1)}$ over ℓ, and observe a fairly steady increase (case (iii)), a fairly steady decrease (case (iv)), or a fairly well defined "v", i.e., decrease and then increase (cases (i) and (ii)) this would provide evidence of a distributional shift.

Since the $Q_m(\ell)$ and $R_m(\ell)$ may be expected to be rather unstable let us average them in blocks of m and define

(12) $$W_1(\ell) = \frac{1}{m} \sum_{i=\ell-m+1}^{\ell} \left(1 - \frac{Q_m(i)}{2}\right)$$

(13) $$W_2(\ell) = \frac{1}{m} \sum_{i=\ell-m+1}^{\ell} \left(1 - \frac{R_m(i)}{2(m-1)}\right).$$

(Note that W_1 and W_2 can not be computed sooner than ℓ=2m).

For $\ell \leq k$, W_1 and W_2 are both unbiased estimators of Σp_j^2 and for $\ell \leq k+2m$, W_1 and W_2 are both unbiased estimators of $\Sigma p_j'^2$. Again in the presence of a distributional change, a perturbation of W_1 and W_2 across k to k+2m should be observed. Although the preceding variance calculations might suggest that $var(W_2) < var(W_1)$ this is not necessarily so since W_1 is an average of independent random variables (see after (10)) while W_2 is an average of dependent variables (see after (11)). While a computation of $var(W_1)$ and $var(W_2)$ could be attempted based on previous calculations the results would be hopeless to compare. However since the $R_m(i)$ are positively correlated one might suspect that the inequality would be reversed. In fact our simulation study in the next section reveals that this is not so, i.e., $var(W_2)$ tends to be much smaller than $var(W_1)$.

In the binomial case with X_ℓ being the result of ℓ^{th} Bernoulli trial we have

$$W_1(\ell) = m - \sum_{i=\ell-m+1}^{\ell} |X_i - X_{i-m}|$$

$$W_2(\ell) = m - \frac{1}{m-1} \sum_{i=\ell-m+1}^{\ell} \left| mX_i - \sum_{r=i-m+1}^{i} X_r \right|.$$

In order to formulate an estimator of k based on either W_1 or W_2, if a disturbance is observed commencing at approximately trial ℓ_o and stabilizing after approximately trial ℓ_o+2m then $\hat{k} = \ell_o$.

3.2 "Departures from Centrality" approaches.

Consider

$$(14)\quad W_3(\ell) = \sum_{j=1}^{r} (S_{m,j}^{(\ell)} - \frac{m}{r})^2 = \sum_{j=1}^{r} (S_{m,j}^{(\ell)})^2 - m^2/r$$

$$(15)\quad W_4(\ell) = \sum_{j=1}^{r} (T_{m,j}^{(\ell)} - \frac{m(m-1)}{2r})^2 = \sum_{j=1}^{r} (T_{m,j}^{(\ell-1)})^2 - \frac{m^2(m-1)^2}{4r}.$$

(Note that W_3 and W_4 can not be computed any sooner than $\ell=m$).

If $\ell \leq k$, W_3 and W_4 indicate "how far" the distribution $\underset{\sim}{p}$ is from the equiprobable cell distribution and if $\ell \geq k+m$ they indicate "how far" $\underset{\sim}{p}'$ is from the equiprobable cell distribution. As ℓ increases from below k to above k+m we ought to be able to observe a change from a fairly stable "distance" to instability and back again to a fairly stable "distance". More precisely

$$E(W_3(\ell)) = \sum_{j=1}^{r} E(S_{m,j}^{(\ell)})^2 - m^2/r$$

$$= \begin{cases} m(m-1)\Sigma p_j^2 + m - m^2/r & \text{if } \ell \leq k \\ (k+m-\ell)(k+m-\ell-1)\Sigma p_j^2 + (\ell-k)(\ell-k-1)\Sigma p_j'^2 & \\ \quad + m + 2(k+m-\ell)(\ell-k)\Sigma p_j p_j' - m^2/r & \\ & \text{if } k+1 \leq \ell < k+m \\ m(m-1)\Sigma p_j'^2 + m - m^2/r & \text{if } \ell \geq k+m. \end{cases}$$

These may be written a bit more suggestively as

$$= \begin{cases} m^2\Sigma(p_j-1/r)^2+m(1-\Sigma p_j^2) & \text{if} \quad \ell \leq k \\ \sum_{j=1}^{r}[(k+m-\ell)(p_j-1/r)+(\ell-k)(p_j'-1/r)]^2 & \\ \quad +(k+m-\ell)(1-\Sigma p_j^2)+(\ell-k)(1-\Sigma p_j'^2) & \\ & \text{if} \quad k+1 \leq \ell < k+m \\ m^2\Sigma(p_j'-1/r)^2+m(1-\Sigma p_j'^2) & \text{if} \quad \ell \geq k+m. \end{cases}$$

While the second set of expressions suggests that departure from the center of the r dimensional simplex is being measured, the first set of expressions reveals that $E(W_3)$ is merely of the form $a(m)\Sigma p_j^2+b(m)$ for $\ell \leq k$ and of the form $a(m)\Sigma p_j'^2+b(m)$ for $\ell \geq k+m$. Thus monitoring W_3 over ℓ is much like studying W_1 or W_2 from the previous section: under a distributional change a departure from relative stability will hopefully be observed across trials k to k+m and then a return to relative stability. All of these remarks are appropriate for W_4. In particular, although the expressions are a bit more complicated it may readily be seen that $E(W_4)$ is of the form $c(m)\Sigma p_j^2+d(m)$ for $\ell \leq k$ and of the form $c(m)\Sigma p_j'^2+d(m)$ for $\ell \geq k+m$.

An argument which lends further support to W_3 is the following. If $\ell \leq k$ then the set $\{S_{m,1}^{(\ell)},\ldots,S_{m,r-1}^{(\ell)}\}$ is a complete and sufficient statistic for the most recent

sample of m observations. It is apparent that

$$H(S^{(\ell)}_{m,\cdot,\ldots,},S^{(\ell)}_{m,r-1}) = \frac{1}{m(m-1)}\sum_{j=1}^{r} S^{(\ell)}_{m,j}(S^{(\ell)}_{m,j} - 1)$$

$= \frac{1}{m(m-1)}\Sigma(S^{(\ell)}_{m,j})^2 - \frac{1}{m-1}$ is the uniformly minimum variance unbiased estimator of Σp_j^2. Similarly, if $\ell \geq k+m$ H is the UMVU of $\Sigma p_j'^2$. Furthermore H and W_3 are linearly related. Thus W_3 will be as effective an index for our purposes as H and may be used equivalently. We comment that since W_1 and W_2 are computed on the basis of a sample of 2m their use is not discouraged by the above.

In concluding this subsection we develop a statistic similar to W_3 and W_4 which is a function of both $S^{(\ell)}_m$ and $T^{(\ell)}_m$. Analogously to prior calculations we have

$$E(S^{(\ell)}_{m,j}) = \begin{cases} mp_j & \text{if} \quad \ell \leq k \\ (m+k-\ell)p_j+(\ell-k)p_j' & \text{if} \quad k+1 \leq \ell < k+m \\ mp_j' & \text{if} \quad k \geq \ell+m \end{cases}$$

and

$$E(T^{(\ell)}_{m,j}) = \begin{cases} \frac{m(m-1)}{2}p_j & \text{if} \quad \ell \leq k \\ \frac{(m+k-\ell)(m+k-\ell-1)}{2}p_j+(\frac{m(m-1)}{2} - \frac{(m+k-\ell)(m+k-\ell-1)}{2}p_j' & \\ & \text{if} \quad k+1 \leq \ell < k+m \\ \frac{m(m-1)}{2}p_j' & \text{if} \quad \ell \geq k+m \end{cases}$$

so that

$$E(T_{m,j}^{(\ell)} - \frac{(m-1)}{2} S_{m,j}^{(\ell)}) = 0 \qquad \text{if} \qquad \ell \leq k,\ \ell \geq k+m$$

$$= \frac{(m+k-\ell)(\ell-k)}{2}(p_j' - p_j)$$

$$\text{if} \qquad k+1 \leq \ell < k+m.$$

Since $\sum_{j=1}^{r} (T_{m,j}^{(\ell)} - \frac{(m-1)}{2} S_{m,j}^{(\ell)}) = 0$ for all ℓ we are led to

$$W_5(\ell) = \sum_{j=1}^{r} (T_{m,j}^{(\ell)} - \frac{(m-1)}{2} S_{m,j}^{(\ell)})^2. \tag{16}$$

(Note that W_5 can't be computed sooner than $\ell=m$.)

The behavior of W_5 should be such that below k and above k+m it will be small and between k and k+m it will tend to be larger.

In the interest of computing $E(W_5)$ we express (16) as

$$W_5(\ell) = \sum_{j=1}^{r} \Big(\sum_{i=\ell-m+1}^{\ell} (\frac{m-1}{2} + i - \ell) X_{ij} \Big)^2.$$

Thus

$$E(W_5(\ell)) = \sum_{j=1}^{r} E\Big[\sum_{i=\ell-m+1}^{\ell} (\frac{m-1}{2} + i - \ell) X_{ij} \Big]^2$$

$$= \sum_{j=1}^{r} \Big[\operatorname{var}\Big(\sum_{i=\ell-m+1}^{\ell} (\frac{m-1}{2} + i - \ell) X_{ij} \Big)$$

$$+ \Big(E\Big(\sum_{i=\ell-m+1}^{\ell} (\frac{m-1}{2} + i - \ell) X_{ij} \Big) \Big)^2 \Big]$$

$$
= \begin{cases}
\sum_{j=1}^{r}\Big[\sum_{i=\ell-m+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)^2 p_j(1-p_j) \\
\qquad +\Big(\sum_{i=\ell-m+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)p_j\Big)^2\Big] & \text{if} \quad \ell \le k \\
\\
\sum_{j=1}^{r}\Big[\sum_{i=\ell-m+1}^{k}\left(\frac{m-1}{2}+i-\ell\right)^2 p_j(1-p_j) \\
\qquad +\sum_{i=k+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)^2 p_j'(1-p_j') \\
\qquad +\Big(\sum_{i=\ell-m+1}^{k}\left(\frac{m-1}{2}+i-\ell\right)p_j+\sum_{i=k+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)p_j'\Big)^2\Big] \\
\qquad \text{if} \quad k+1 \le \ell < k+m \\
\\
\sum_{j=1}^{r}\Big[\sum_{i=\ell-m+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)^2 p_j'(1-p_j') \\
\qquad +\Big(\sum_{i=\ell-m+1}^{\ell}\left(\frac{m-1}{2}+i-\ell\right)p_j'\Big)^2\Big] & \text{if} \quad \ell \ge k+m
\end{cases}
$$

Although these expressions may be simplified somewhat it is more crucial to note that again $E(W_5)$ is of the form $e(m)\Sigma p_j^2+f(m)$ for $\ell \le k$ and of the form $e(m)\Sigma p_j'^2+f(m)$ for $\ell \ge k+m$. Thus W_5 may be monitored over ℓ as our intuition suggests and analogous to W_3 and W_4.

In the binomial case our expressions for W_3, W_4 and W_5 reduce to

$$W_3(\ell) = 2\left(\sum_{i=\ell-m+1}^{\ell} X_i - \frac{m}{2}\right)^2$$

$$W_4(\ell) = 2\left(\sum_{i=\ell-m+1}^{\ell-1} (m+i-\ell)X_{i+1} - \frac{m(m-1)}{4}\right)^2$$

$$W_5(\ell) = 2\left(\sum_{i=\ell-m+1}^{\ell} \left(\frac{m-1}{2}+i-\ell\right)X_i\right)^2.$$

As at the end of section 3.1, $\hat{k}=\ell_o$ provides an estimator of k.

3.3 "Test of Homogeneity" approaches.

A test of homogeneity seems to quite naturally suggest itself for this sort of problem. We are trying to detect for a collection of observations whether one multinomial distribution is generating them or whether two different multinomials are generating them. Discovering which of these two hypotheses is true is precisely the raison d'etre for chi-square based test of homogeneity statistics.

However in our present circumstance where we have a continuing sequence of observations there is certainly no unique choice of partitioning into groups to establish comparisons. However since we postulate but one distributional change across any given portion of the sequence

we are examining, a partition into more than two groups seems likely to obscure or confound the detection problem.

If we confine ourselves to two blocks of observations, as an initial attempt, we would compare the first m observations to the next m observations by computing the usual χ^2 statistic and continue successively comparing the i^{th} block of m observations with the $i+1^{st}$ block, i=2,3,.... Our statistic depends on i and m and in the notation developed thus far becomes

$$V_1(i,m) = \sum_{j=1}^{r} \frac{(S_{m,j}^{((i+1)m)} - S_{m,j}^{(im)})^2}{S_{m,j}^{((i+1)m)} + S_{m,j}^{(im)}}. \tag{17}$$

In monitoring V_1, if no shift occurred during trials im+1 to i(m+1) we expect V_1 small, otherwise it should be large. As we have indicated previously the selection of m depends on r. It should be at least 5r to discourage empty cell problems. More precisely we would like the expected cell frequencies to be at least three to five. With the true distributions unknown we instead suggest $m \cdot \frac{1}{r}$ at least five. Since we are not concerned with the distribution of our monitoring index we might drop the denominator from V_1 and simply examine

$$V_2(i,m) = \sum_{j=1}^{r} (S_{m,j}^{((i+1)m)} - S_{m,j}^{(im)})^2. \tag{18}$$

Another strategy which is an adaptive modification of V_1 is the following. If $V_1(1,m)$ is small pool the first 2m observations into one sample and compare this group with the third block of m observations. More generally this idea leads to a comparison of the first im observations with the next set of m observations via

$$(19) \qquad V_3(i,m) = \sum_{j=1}^{r} \frac{(S_{im,j}^{(im)} - i\, S_{m,j}^{((i+1)m)})^2}{i(S_{im,j}^{(im)} + S_{m,j}^{((i+1)m)})}$$

if $V_3(1,m)$, $V_3(2,m),\ldots,V_3(i-1,m)$ are small. Again the denominator in (19) may be suppressed using instead

$$(20) \qquad V_4(i,m) = \sum_{j=1}^{r} (S_{im,j}^{(im)} - i\, S_{m,j}^{((i+1)m)})^2.$$

All of these statistics suffer one drawback. When a distributional change occurs it will occur somewhere within a block of m trials. When our statistics are computed across this change we will be comparing m or more observations from the old distribution with m observations some from the old and some from the new. But the most dramatic difference should be revealed if we compare the m trials just prior to a shift with the m trials just after a shift. Moreover we would prefer a statistic which can be computed at successive trials

(such as those of the previous sections) rather than at every m^{th} trial. A statistic which accomplishes both these objectives is

$$W_6(\ell) = \sum_{j=1}^{r} \frac{(S_{m,j}^{(\ell)} - S_{m,j}^{(\ell-m)})^2}{(S_{m,j}^{(\ell)} + S_{m,j}^{(\ell-m)})} \tag{21}$$

or deleting the denominator

$$W_7(\ell) = \sum_{j=1}^{r} (S_{m,j}^{(\ell)} - S_{m,j}^{(\ell-m)})^2. \tag{22}$$

(Note that W_6 and W_7 can not be computed any sooner than $\ell=2m$.)

In observing W_6 and W_7 over ℓ we expect them to begin to increase at $\ell=k+1$ peaking near $\ell=k+m$ and then decreasing until $\ell=k+2m$. The occurrence of a spike of this sort suggests that a distributional change has occurred.

Although $E(W_6)$ is rather hopeless to compute, $E(W_7)$ is straightforwardly obtained directly from results after (15), i.e.,

$$E(W_7(\ell)) = \sum_{j=1}^{r} E(S_{m,j}^{(\ell)} - S_{m,j}^{(\ell-m)})^2$$

$$= \sum_{j=1}^{r} [E(S_{m,j}^{(\ell)})^2 + E(S_{m,j}^{(\ell-m)})^2 - 2E(S_{m,j}^{(\ell)})E(S_{m,j}^{(\ell-m)})]$$

$$= \begin{cases} 2m(m-1)\Sigma p_j^2 + 2m - 2m^2\Sigma p_j^2 & \text{if } \ell \le k \\ (k+m-\ell)(k+m-\ell-1)\Sigma p_j^2 + (\ell-k)(\ell-k-1)\Sigma p_j'^2 & \\ \quad +2(k+m-\ell)(\ell-k)\Sigma p_j p_j' + m(m-1)\Sigma p_j^2 & \\ \quad +2m-2m(m+k-\ell)\Sigma p_j^2 - 2m(\ell-k)\Sigma p_j p_j' & \\ & \text{if } k+1 \le \ell < k+m \\ (k+m-\ell)(k+m-\ell-1)\Sigma p_j^2 + (\ell-k)(\ell-k-1)\Sigma p_j'^2 & \\ \quad +2(k+m-\ell)(\ell-k)\Sigma p_j p_j' + m(m-1)\Sigma p_j'^2 & \\ \quad +2m-2m(m+k-\ell)\Sigma p_j p_j' - 2m(\ell-k)\Sigma p_j'^2 & \\ & \text{if } k+m \le \ell < k+2m \\ 2m(m-1)\Sigma p_j'^2 + 2m - 2m^2\Sigma p_j'^2 & \text{if } \ell \ge k+2m. \end{cases}$$

After simplification these expressions become

$$\begin{cases} 2m(1-p_j^2) & \text{if } \ell \le k \\ (\ell-k)^2\Sigma(p_j-p_j')^2 + (\ell-k)(\Sigma p_j^2 - \Sigma p_j'^2) + 2m(1-\Sigma p_j^2) & \\ & \text{if } k+1 \le \ell < k+m \\ (k+m-\ell)^2\Sigma(p_j-p_j')^2 + (k+m-\ell)(\Sigma p_j'^2 - \Sigma p_j^2) + 2m(1-\Sigma p_j'^2) & \\ & \text{if } k+m \le \ell < k+2m \\ 2m(1-\Sigma p_j'^2) & \text{if } \ell > k+2m. \end{cases}$$

Thus as with our previous W statistics, $E(W_7)$ depends on $\underset{\sim}{p}$ and $\underset{\sim}{p}'$ only through Σp_j^2, $\Sigma p_j p_j'$, and $\Sigma p_j'^2$. Moreover $E(W_7)$ depends only on Σp_j^2 if $\ell \leq k$, only on $\Sigma p_j'^2$ if $\ell \geq k+2m$. $E(W_7)$ also shows that W_7 may be expected to increase and then decrease across the trials k to k+2m.

In the binomial case W_6 and W_7 reduce to

$$W_6(\ell) = \frac{2m\left(\sum_{i=\ell-m+1}^{\ell} X_i - \sum_{i=\ell-2m+1}^{\ell-m} X_i\right)^2}{\left(2m - \sum_{i=\ell-2m+1}^{\ell} X_i\right)\left(\sum_{i=\ell-2m+1}^{\ell} X_i\right)}$$

$$W_7(\ell) = 2\left(\sum_{i=\ell-m+1}^{\ell} X_i - \sum_{i=\ell-2m+1}^{\ell-m} X_i\right)^2.$$

As in the previous sections $\hat{k}=\ell_o$ provides an estimator of k.

4. The Results of a Simulation Study.

An extensive simulation study was undertaken to compare the performance of the 7 W and 4 V statistics under a variety of circumstances. Table 1 enumerates 10 of the more interesting cases examined. A total of 300 trials were run for each case with k=150 and blocks, m, of 20, 30, 40 and 50. The r values 3, 5, and 7 were selected as the likely range of application. The distributions were selected such that for some, the change should be easy to

discern while for others, it should be more difficult. They also reflect an assortment of different combinations for Σp_j^2, $\Sigma p_j p_j'$, and $\Sigma p_j'^2$.

We will comment on each case briefly but first some general observations may be made.

(i) The cases revealed the crucial dependence of the selection of m based on r. With r=3, m=20 or 30 were effective while with r=7 m at least 40 and usually 50 was necessary. A practical rule of thumb would be to take m <u>at least</u> 7 to 10 times r if feasible.

(ii) In all cases studied var(W_3) was much smaller than var(W_4) making $W_3(\ell)$ far more effective than $W_4(\ell)$ in revealing a pattern of distributional shift as opposed to random variation.

(iii) Similarly in all cases studied var(W_2) was much smaller than var(W_1) making $W_2(\ell)$ more effective than $W_1(\ell)$.

(iv) None of the V statistics worked as well as either $W_6(\ell)$ or $W_7(\ell)$ so that they will be dispensed with in the remaining discussion.

(v) $W_6(\ell)$ and $W_7(\ell)$ never failed to respond to a distributional shift but occasionally (in rather erratic data sequences) indicated a false shift.

(vi) Because the W_3, W_4 and W_5 statistics will often be of the order of 10^3 or 10^4 logarithms were taken to give a more tractable scale and to more clearly reveal patterns.

Case #	r	Distribution before change Distribution after change	Σp_j^2, $\Sigma p_j p_j'$, $\Sigma p_j'^2$
1	3	(.6,.3,.1) (.1,.6,.3)	.46, .27, .46
2	3	(.5,.3,.2) (.8,.1,.1)	.38, .45, .66
3	3	(.3,.3,.4) (.6,.3,.1)	.34, .31, .46
4	3	(.4,.4,.2) (.4,.2,.4)	.36, .32, .36
5	5	(.5,.2,.1,.1,.1) (.1,.5,.2,.1,.1)	.32, .19, .32
6	5	(.3,.2,.2,.1,.1) (.6,.2,.1,.05,.05)	.22, .255, .415
7	5	(.2,.2,.2,.2,.2) (.4,.3,.1,.1,.1)	.20, .20, .28
8	5	(.2,.3,.1,.2,.2) (.2,.1,.3,.2,.2)	.22, .18, .22
9	7	(.4,.1,.1,.1,.1,.1,.1) (.1,.4,.1,.1,.1,.1,.1)	.22, .13, .22
10	7	(.2,.2,.2,.1,.1,.1,.1) (.3,.3,.2,.05,.05,.05,.05)	.16, .18, .23

Table 1: A Sampling of Simulation Cases Studied

(vii) $W_5(\ell)$ consistently revealed multiple spikes over ℓ thereby clouding the perception of a distributional change. However in virtually every case considered its absolute peak was within 5 trials of k+m/2.

We now turn to the individual cases.

Case 1: This should be an easy shift to detect. Nonetheless W_1 and W_2 were ineffective until m=40 and were only really clear at m=50. W_3 and W_4 were reasonably clear throughout. W_5 developed a more sharply defined unique peak with increasing m. W_6 and W_7 were excellent even at m=20.

Case 2: Again this should be an easy shift to detect. W_1 and W_2 were both fairly good at m=20, 30. By m=40, 50, W_2 was excellent and W_1 quite good. W_3 and W_4 were ineffective until m=40. W_5 was erratic but absolute peaks were always correct. W_7 and to a lesser extent W_6 were bothered by early idiosyncrasies in the sequence which smoothed out by m=40.

Case 3: This shift should be a bit more difficult to detect. W_1 was ineffective even at m=50 although W_2 was good by m=40. W_3 was better than W_4 although neither was really sharp until m=50. W_5 was fairly clear throughout. W_6 and W_7 were both good even at m=20.

Case 4: This is clearly the most difficult to detect of the three-cell cases presented. W_1 was ineffective throughout. W_2 was only effective at m=50. W_3 and W_4 were quite good particularly by m=40. W_5 was too erratic to be helpful. W_6 and W_7 were also ineffective for each m.

Case 5: This shift, analogous to case 1, should be easy to detect. W_1 was not effective until m=50. W_2 was better being reasonably clear by m=40. W_3 was good at m=40 and excellent at m=50 while W_4 never responded. W_5 was fairly clear throughout. W_6 and W_7 were excellent surprisingly even at m=20.

Case 6: This shift analogous to case 2 should also be easy to detect. W_2 was excellent throughout with W_1 effective at m=40. W_3 and W_4 were also ineffective until m=40. W_5 was too erratic to be helpful. W_6 was fairly clear at m=30 and quite good at m=40, 50, while W_7 did not become clear until 50. As in case 2 both W_6 and W_7 seemed somewhat bothered by early idiosyncrasies in the sequence.

Case 7: This case should be more difficult to detect than the previous two. W_2 is excellent by m=40 with W_1 clear by m=50. W_3 is sharper than W_4 but both are good by m=40. W_5 is fairly clear throughout. W_6 and W_7 both detected a false

shift at ℓ=100 which smoothed out somewhat but not completely by m=50.

Case 8: This case analogous to case 4 is clearly the most difficult of the five cell cases presented. W_1 never responded while W_2 was quite good at m=50. W_4 was erratic throughout while W_3 was marginally effective by m=40. W_5 was quite good throughout giving better performance relative to the other statistics than in previous cases. W_6 and W_7 were good particularly at m=40, 50.

Case 9: With 7 cells, m=20 and m=30 ought not work well for any of the W's. Nonetheless W_6 and W_7 were excellent by m=30. W_1 was poor throughout with W_2 finally becoming reasonably clear by m=50. W_3 and W_4 were poor throughout while W_5 was fairly good throughout.

Case 10: This case should be much more difficult than Case 9. Nonetheless W_1 and W_2 were both quite good by m=40. Also W_3 and W_4 were effective after m=40. W_5 was too erratic to be effective. W_7 was poor throughout with W_6 a bit better by m=50.

In summary from the methods of section 3.1, $W_2(\ell)$ is clearly the better choice. From the methods of section 3.2, $W_3(\ell)$ is the best choice and from the methods of section 3.3, $W_6(\ell)$ is the best choice. Amongst W_2, W_3, and W_6 it is not possible to select an overall best

choice. All three should be effective with m large and all three can be monitored concurrently to mutually confirm a detected distribution change.

Acknowledgments: I would like to thank Herbert Solomon for bringing this problem to my attention and John Mahoney for his development of the computer program used in the simulation study.

References

1. Chernoff, H. and Zacks, S. (1964) "Estimating the Current Mean of A Normal Distribution which is Subjected to Changes in Time", Annals of Math. Stat. 35, p. 999-1018.

2. Kander, Z. and Zacks, S. (1961) "Test Procedures for Possible Changes in Parameters of Statistical Distributions Occurring at Unknown Time Points", Annals of Math. Stat. 37, p. 1196-1210.

3. Page, E.S. (1955) "A Test for a Change in A Parameter Occurring at an Unknown Point", Biometrika 42, p. 523-526.

4. Sclove, S.L. (1968) "Remarks on the Problem of 'Homogenization of Bernoulli Trials'", Stanford Technical Report No. 137.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: TR-282	2. GOVT ACCESSION NO.: AD-A087898
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Detecting Change Points in Sampling from Multinomial Distributions,	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Alan E. Gelfand	8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0475; DAAG29-77-G-0031
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Stanford University, Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Statistics and Probability Program Code 436, Arlington, VA 22217	12. REPORT DATE: March, 1980; 13. NUMBER OF PAGES: 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED; 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution is Unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This report partially supported under U.S. Army Research Office Grant DAAG29-77-G-0031.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multinomial distribution, change point, weighted sums, unweighted sums, central limit theory approaches, departure from centrality approaches, test of homogeneity approaches.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Please see reverse side.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

332580

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT NO. 282

DETECTING CHANGE POINTS IN SAMPLING FROM MULTINOMIAL DISTRIBUTIONS

The question of if and when a change in the underlying cell frequencies has occurred during the observation of a sequence of categorical variables is analyzed. Several techniques are developed that may be loosely grouped into three classifications: (i) Central Limit Theorem approaches, (ii) Departure from Centrality approaches and (iii) Test of Homogeneity approaches. The approaches monitor the sequence either at every trial or at every k-th trial. All approaches construct statistics which are functions of sequential sums either weighted or unweighted. The behavior of these sums and the statistics developed from them are discussed in detail. A large scale simulation study is discussed in an attempt to assess the performance of the approaches. Three approaches emerge as most promising but a clear choice among these is not possible at present.

S N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)